Modélisation statistique d'un procédé de centrifugation

Zhanhao Liu[†,‡,1], Marion Perrodin[‡], Thomas Chambrion[†], Radu S. Stoica [†],

[†] *Université de Lorraine, CNRS, IECL, F-54000 Nancy, France*
[‡] *Saint-Gobain Research Paris, 39 Quai Lucien Lefranc, 93300 Aubervilliers*

Résumé. Cet article présente une analyse statistique des données issues d'un procédé de centrifugation utilisé à Saint-Gobain. Les différentes corrélations entre les variables enregistrées ont été analysées via une ACP, et sur cette base plusieurs modèles statistiques ont été proposés. L'objectif final est de proposer un processus de contrôle de procédé de centrifugation à travers cette analyse statistique. Ce travail est actuellement en cours mais les résultats obtenus indiquent déjà quelles étapes du procédé industriel pourraient jouer de manière prépondérante sur la qualité du produit final.

Mots-clés. Analyse en Composantes Principales, modélisation linéaire, validation croisée, Industrie 4.0

Abstract. This article describes a statistical analysis applied on a centrifugation process of Saint-Gobain. The correlations within data were analysed by PCA, and some statistical models were built using information from the previous step. The final purpose is to build a control law for the centrifugation process through this statistical analysis. The work is still ongoing, but the first results already show which process steps play an important role in the final product's quality.

Keywords. Principal Component Analysis, linear model, cross-validation, Industry 4.0

1 Introduction

Les lignes de production de Saint-Gobain sont intrinsèquement complexes car composées de multiples machines effectuant chacune une étape du processus. Ces machines doivent interagir et collaborer afin d'atteindre le cahier des charges demandé. La complexité de chaque étape, les nombreuses interdépendances et tous les facteurs déterminant le fonctionnement de la ligne font que la construction d'un modèle physique d'une telle ligne de production est extrêmement compliquée. Dans ce contexte, l'analyse statistique nous semble un outil prometteur pour l'apport d'informations mettant en évidence les éléments clefs dans le processus industriel étudié, d'autant plus que les lignes de production sont

[1]Email de contact : zhanhao.liu@univ-lorraine.fr

de plus en plus instrumentées, et remontent de grandes quantités de données. En plus de l'étude et de la compréhension d'un phénomène physique difficilement modélisable, nous souhaitons utiliser l'analyse statistique pour alimenter l'approche *Industrie 4.0*. Celle-ci consiste en la mise en place d'outils, issus principalement du traitement des données remontées sur la ligne de production, aidant les opérateurs humains à prendre des décisions tout en optimisant énergétiquement et qualitativement la production.

Nous développons ici une telle approche pour un procédé de fabrication de tuyaux en fonte utilisé par Saint-Gobain Pont-à-Mousson, dont voici les principales étapes :

- le *basculement* : de la fonte en fusion est versée dans un récipient, que nous appelons un "basket", contenant l'équivalent en fonte de plusieurs tuyaux. Le moule est ensuite alimenté en fonte par des inclinaisons successives du basket.

- la *translation* : un chariot porte le moule rempli de fonte, en faisant des aller-retour sur une rampe entre un point haut et un point bas. Au point bas, qui est aussi le point de repos, la fonte est déjà transformée en tuyau, et le tuyau est livré. Le chariot remonte à vide pour être rempli à nouveau.

- la *rotation* : la rotation du moule permet la formation des tuyaux par force centrifuge. La rotation démarre en même temps que le versement de la fonte, et elle continue pendant la translation.

- l'*extraction* : quand le chariot est au point bas, un bras mécanique fait sortir le tuyau du moule.

Une mesure d'efficacité de ce procédé de fabrication de tuyaux est la différence, notée y, entre la masse du tuyau réalisée et la masse demandée par le cahier des charges. L'objectif de cette étude est de comprendre le rôle de chaque étape dans l'obtention d'une production de qualité, soit minimiser y sous la contrainte $y \geq 0$. Nous allons étudier les différentes corrélations existantes entre les données enregistrées lors du procédé de fabrication, puis nous proposerons des modèles statistiques pour l'estimation de l'écart de masse y. Dans la suite de cet article, nous allons tout d'abord présenter les données enregistrées lors du procédé de fabrication, puis les résultats d'une analyse en composantes principales. Ensuite, plusieurs modèles seront proposés et étudiés, parmi lesquels des modèles linéaires et des modèles de convolution avec des dépendances "spatio-temporelles". Enfin, nous présenterons les conclusions tirées de ces premiers résultats et les perspectives qui en découlent.

2 Données

Les données sont réparties dans plusieurs fichiers, chaque fichier contenant les informations des tuyaux fabriqués durant un poste de 8 heures.

2.1 Variables

Différentes variables sont mesurées pendant les étapes de fabrication. Notamment, sont mesurés durant le basculement, les mouvements angulaires du basket, les durées d'action et les vitesses de basculement ; durant la translation, les distances parcourues par le chariot et les durées ; durant la rotation, les vitesses de rotation du moule et les durées des "régimes" ; et enfin durant l'extraction, les durées de l'action.

Des variables quadratiques sont créées directement dans la base pour caractériser la quantité de fonte versée : ce sont les vitesses de basculement multipliées par les durées appliquées. Quant aux variables qualitatives, il y a par exemple des indicateurs d'anomalies ou la numérotation des tuyaux. Au total, nous avons initialement 61 variables quantitatives et 28 variables qualitatives, qui ont été standardisées avant leur utilisation.

2.2 Individus

Nous désignerons désormais un tuyau fabriqué par le terme "individu". Le terme "basket" désigne à la fois l'outil de versement de fonte et les tuyaux réalisés avec la fonte contenue par celui-ci entre deux remplissages. Par exemple, ce que nous appellerons le 4ième basket correspondra aux tuyaux réalisés entre le 4ième et le 5ième remplissage du basket proprement-dit au cours d'un même poste. Le numéro de tranche est quant à lui défini par rapport au niveau de fonte restante dans le basket lorsque la fabrication du tuyau concerné commence. Le nombre de tuyaux réalisés avec un basket n'est cependant pas constant car le remplissage de celui-ci peut fluctuer et certaines tranches peuvent ne pas être présentes pour certains baskets.

Les individus sont caractérisés par 3 facteurs : le poste, le basket et le numéro de tranche. On note le tuyau du $i$ième basket, de la $j$ième tranche et du $k$ième poste comme $T_{i,j,k}$. A chaque $T_{i,j,k}$ correspondent donc des mesures ou des paramètres $X_{i,j,k}$ et une cible $y_{i,j,k}$.

Nous avions initialement 50225 individus dans notre base de données. Nous avons sélectionné aléatoirement 50% des postes pour le training set, 25% pour le cross-validation set, et 25% pour le test set.

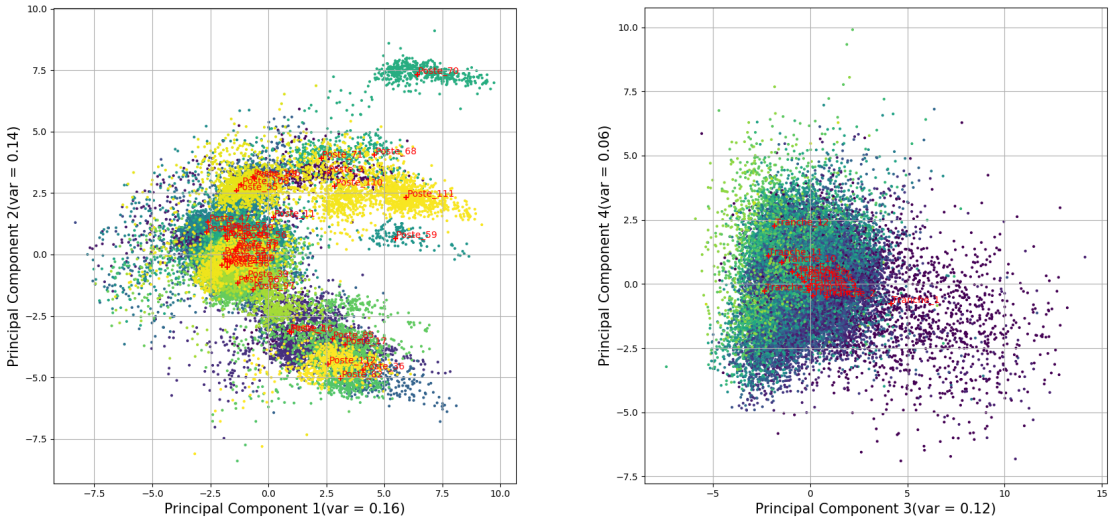
Nous avons considéré les individus ayant une variable dont la valeur standardisée absolue était supérieure à 5 comme des outliers. Les outliers et les individus avec un indicateur d'anomalie actif ont été supprimés de nos différents data sets. Après filtrage, nous avons 21149 individus dans le training set, 14562 dans le cross-validation set et 9611 dans le test set.

3 Analyse exploratoire

Les données sont traitées via une analyse en composantes principales (F. Husson et al. (2017)), qui nous indique qu'en gardant 19 composantes, on préserve 90% de variation

globale des données, tout en réduisant la dimension du problème.

Les cercles de corrélation nous montrent des corrélations significatives entre certaines variables caractérisant la même étape du processus. Cela indique que ces variables pourraient être regroupées pour la construction d'un modèle. Toutefois, le fait que la variation de la cible soit faiblement caractérisée par cette nouvelle représentation des données, nous donne de faibles espoirs pour qu'un modèle simple puisse expliquer nos données. Cette hypothèse sera d'ailleurs rejetée dans la section suivante.



(a) plan factoriel

(b) plan des composantes 3 et 4

Figure 1: Projection des individus. (a) : chaque couleur correspond à un poste. (b) : chaque couleur correspond à une tranche de basket. Plus la couleur est claire, plus le numéro de tranche est grand.

Les projections des individus dans les plans des quatre premières composantes sont présentées dans la Figure 1. Nous observons que les individus forment des agrégats en fonction de leur appartenance à une tranche et/ou à un poste. Un test statistique bootstrap valide d'ailleurs cette affirmation. Cela suggère une dépendance temporelle (dans un poste) et spatiale (entre les mêmes tranches) entre les individus. La dépendance temporelle est due aux différentes conditions de travail : la température ambiante, la composition de la fonte, le type de basket utilisé, les habitudes des opérateurs ; tandis que la dépendance spatiale est liée à l'angle de versement du début cycle. Bien que l'information concernant le type de basket ne soit pas disponible sur notre base de données, nous avons pu mettre en évidence son importance à travers des méthodes de classification. Il nous apparaît important que notre modèle final puisse aussi expliquer la répartition en agrégats des individus.

4 Modélisation

Cette section présente plusieurs étapes, de la plus simple à la plus élaborée, pour la construction des modèles qui pourraient expliquer nos données. Ce travail est en cours.

Suite à nos observations précédentes, les premières tranches ont été supprimées des trois sets du fait de leur comportement spécifique, qui aurait pu interférer dans nos modélisations.

4.1 Régression simple

L'hypothèse la plus simple à vérifier concernant la variation de la cible est la suivante : peut-on l'expliquer par la combinaison linéaire des données observées lors de son processus de fabrication, du basculement jusqu'à l'extraction ? Avec les notations introduites, un tel modèle peut s'écrire comme :

$$\hat{y}_{i,j,k} = \theta_0 + \sum_{l=1}^{p} \theta_l x_{i,j,k,l} \tag{1}$$

avec le vecteur des paramètres $\Theta = (\theta_0, ..., \theta_p)$, les variables explicatives $x_{i,j,k,l}$ avec $l = 1, ..., p$ et $\hat{y}_{i,j,k}$ la cible du tuyau $T_{i,j,k}$.

Suite à l'ACP, un regroupement des variables est effectué. Ceci nous permet de réaliser sur ces données une régression linéaire avec $p = 76$ paramètres. Après plusieurs expériences, la méthode d'estimation Lasso (Tibshirani (1996)) nous indiquait 28 paramètres significativement différents de zéro, alors que le meilleur score R^2 obtenu était seulement de 0.1. Ce faible score était attendu suite à l'analyse ACP. Il nous a permis d'écarter l'hypothèse la plus simple. Dans ce contexte, nous avons décidé d'introduire plus de dépendances dans le modèle (T. Hastie et al. (2017)).

4.2 Introduction de dépendances supplémentaires

La significativité de certains coefficients dans le modèle de régression linéaire simple indique que la variation de y du tuyau $T_{i,j}$ n'est pas indépendante des étapes de fabrication du tuyau lui-même. Le modèle que l'on propose ici introduit, en plus, des dépendances relatives à la fabrication des autres tuyaux sur le même poste $T_{i,j-1}$, $T_{i,j-1}$.

Dans ce contexte nous proposons d'estimer la variation du y du tuyau $T_{i,j}$ par :

$$\hat{y}_{i,j} = \theta_0 + \sum_{l=1}^{p} (\theta_l x_{i,j,l} + \theta_{l+p} x_{i-1,j,l} + \theta_{l+2p} x_{i,j-1,l}) \tag{2}$$

avec $\Theta = (\theta_0, ...\theta_{3p})$ les paramètres du modèle et x les mesures associées pour chaque tuyau considéré. Le modèle estimé par Lasso a un meilleur score R^2 de 0.196 avec 78

variables actives. Cela indique un meilleur comportement de ce modèle par rapport au modèle linéaire simple, et valide l'idée d'étude des dépendances entre tuyaux.

Clairement, le modèle doit encore être amélioré. Nous envisageons des modèles "spatiaux" de la forme :

$$\hat{y}_{i,j} = \theta_0 + \theta_1 y_{i-1,j} + \theta_2 y_{i,j-1} + \sum_{l=1}^{p} (\theta_{l+2} x_{i,j,l} + \theta_{l+2+p} x_{i-1,j,l} + \theta_{l+2+2p} x_{i,j-1,l}) \qquad (3)$$

comparables à ceux de (Antoniadis et al., 1992), (Gaetan et Guyon, 2009) et (Cressie, 2015).

5 Conclusion

Nous avons appliqué une analyse statistique des données sur le procédé de centrifugation de Saint-Gobain. Les résultats de l'analyse en composantes principales montrent que l'écart entre la masse réalisée du tuyau et sa consigne est peu corrélé avec les variables enregistrées dans chaque étape de fabrication. Le meilleur modèle linéaire obtenu a un score R^2 de 0.196 avec 78 variables actives. Ce modèle est linéaire avec quelques variables quadratiques.

Nous avons constaté une dépendance temporelle et spatiale entre les tuyaux réalisés, et en conséquence, une méthode prenant en compte les tuyaux réalisés précédemment sera étudiée. A partir du meilleur modèle obtenu, une loi de contrôle sera construite afin de minimiser l'écart entre la masse réalisée du tuyau et sa consigne.

Bibliographie

A. Antoniadis, J. Berruyeur et R. Carmona (1992), *Regression non lineaire et applications.*
N. Cressie (2015), *Statistics for Spatial Data.*
C. Gaetan et X. Guyon (2009), *Spatial Statistics and Modeling.*
F. Husson, S. Lê et J. Pagès (2017), *Exploratory Multivariate Analysis by Example Using R.*
T. Hastie, R. Tibshirani et J. Friedman (2017), *The Elements of Statistical Learning*
R. Tibshirani (1996), *Regression shrinkage and selection via the Lasso*, Journal of the Royal Statistical Society.